RefSeq Functional Elements: An Annotated Dataset of Validated Non-Genic Elements for Variant Interpretation and Functional Discovery in the Human Genome

Catherine M. Farrell, Terence D. Murphy, and the RefSeq Curation and Development Team*
National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Follow NCBI on:

Abstract: The human genome contains many non-genic elements that play roles in gene regulation, chromosome organization, recombination, repair or DNA replication. Human disease can result from sequence variation in those elements, with many genome-wide association studies indicating disease-associated variation in non-coding regions. The locations of gene regulatory elements can be predicted from several large-scale epigenomic mapping projects, but those data are not generally visible in traditional genome annotation, are difficult to interpret in the absence of specialized research knowledge or customized displays, and do not always show function when tested experimentally. NCBI has therefore introduced a more accessible dataset, RefSeq Functional Elements (www.ncbi.nlm.nih.gov/refseq/functionalelements/), which are annotated on the human genome alongside conventional genes. This curated dataset, which is restricted to known elements from published experimental data, includes richly annotated RefSeq records and accompanying descriptive records in the Gene database (www.ncbi.nlm.nih.gov/gene/). The dataset includes known enhancers, silencers, recombination regions, and other non-genic regions with experimentally-validated function. As of NCBI's Updated Annotation Release 109.20190905 on the GRCh38.p13 genome assembly, the dataset includes over 3.8K GeneIDs and 8.9K feature annotations, with further growth expected for future NCBI annotation releases. The dataset is publicly available for FTP download (ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/). Feature annotation can be visualized in the 'Biological regions' track available in NCBI browsers, including the Genome Data Viewer. This data track is particularly useful when viewed alongside other available tracks, such as variation tracks from dbSNP, ClinVar or dbVar, or study-specific custom tracks or track hubs. These non-genic annotations provide insights into non-coding genome function. They are valuable for basic discovery of gene regulatory regions, interpretation of non-coding variants, or as known positive controls for genome-wide studies aimed at discovering additional elements.

RefSeq Functional Elements definition and scope

1

Definition:

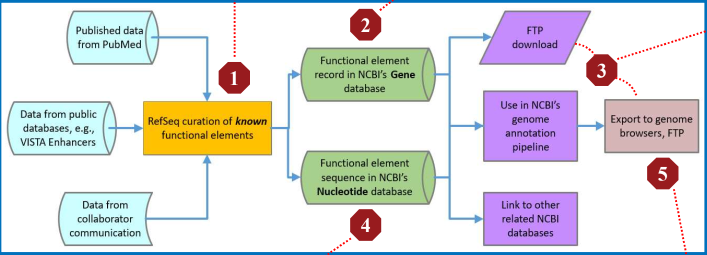
- Any **non-genic** genomic element that has functional significance **based on experimental support**, and is not otherwise considered a conventional gene

Types:

- **Gene regulatory elements**, e.g. enhancers, silencers, promoters, protein binding sites
- **Known structural elements**, e.g. boundary elements, matrix/scaffold-associated regions, other structural regions associated with chromatin conformation
- **Other elements of functional importance**, e.g. clinically-significant recombination hotspots, well-defined replication origins

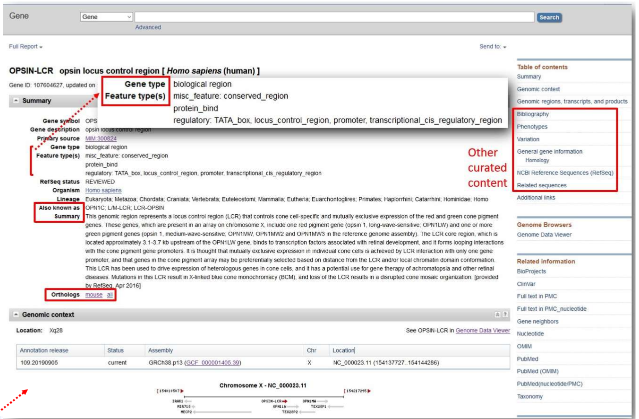
Annotation scope:

- Functional elements that have been **experimentally validated**
- **Human** and **mouse** elements
- Priority for **elements implicated in human disease**


Published data from PubMed
Data from public databases, e.g., VISTA Enhancers
Data from collaborator communication
1
RefSeq curation of known functional elements
2
Functional element record in NCBI's Gene database
Functional element sequence in NCBI's Nucleotide database
4
FTP download
3
Use in NCBI's genome annotation pipeline
Link to other related NCBI databases
Export to genome browsers, FTP
5

Gene database records (www.ncbi.nlm.nih.gov/gene/)

2



Gene | Gene | Search | Advanced

Full Report | Send to:

OPSIN-LCR opsin locus control region [*Homo sapiens* (human)]

Gene ID: 107604627, updated on

Summary

Gene type	biological region
Feature type(s)	misc_feature: conserved_region protein_bind regulatory: TATA_box, locus_control_region, promoter, transcriptional_cis_regulatory_region

Gene symbol	OPS
Gene description	opsin locus control region
Primary source	MIM:300824
Gene type	biological region
Feature type(s)	misc_feature: conserved_region protein_bind regulatory: TATA_box, locus_control_region, promoter, transcriptional_cis_regulatory_region
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	OPN1C; L/M-LCR; LCR-OPSIN
Summary	This genomic region represents a locus control region (LCR) that controls cone cell-specific and mutually exclusive expression of the red and green cone pigment genes. These genes, which are present in an array on chromosome X, include one red pigment gene (opsin 1, long-wave-sensitive; OPN1LW) and one or more green pigment genes (opsin 1, medium-wave-sensitive; OPN1MW, OPN1MW2 and OPN1MW3 in the reference genome assembly). The LCR core region, which is located approximately 3.1-3.7 kb upstream of the OPN1LW gene, binds to transcription factors associated with retinal development, and it forms looping interactions with the cone pigment gene promoters. It is thought that mutually exclusive expression in individual cone cells is achieved by LCR interaction with only one gene promoter, and that genes in the cone pigment array may be preferentially selected based on distance from the LCR and/or local chromatin domain conformation. This LCR has been used to drive expression of heterologous genes in cone cells, and it has a potential use for gene therapy of achromatopsia and other retinal diseases. Mutations in this LCR result in X-linked blue cone monochromacy (BCM), and loss of the LCR results in a disrupted cone mosaic organization. [provided by RefSeq, Apr 2016]
Orthologs	mouse all

Other curated content

Genomic context

Location: Xq28 | See OPSIN-LCR in Genome Data Viewer

Annotation release	Status	Assembly	Chr	Location
109.20190905	current	GRCh38.p13 (GCF_000001405.39)	X	NC_000023.11 (154137727..154144286)

Chromosome X - NC_000023.11

Table of contents: Summary; Genomic context; Genomic regions, transcripts, and products; Bibliography; Phenotypes; Variation; General gene information; Homology; NCBI Reference Sequences (RefSeq); Related sequences; Additional links

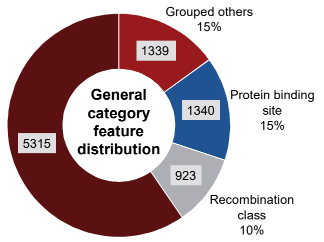
Genome Browsers: Genome Data Viewer

Related information: BioProjects; ClinVar; Full text in PMC; Full text in PMC_nucleotide; Gene neighbors; Nucleotide; OMIM; PubMed; PubMed (OMIM); PubMed(nucleotide/PMC); Taxonomy

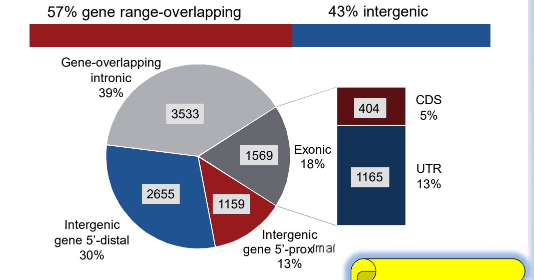
Analysis of RefSeq Functional Elements

From NCBI Annotation Release 109.20190905:

- Total feature count: 8,917
- GeneID count: 3,775
- Genome coverage: 5.6 Mb



Feature locations relative to genes



Overlapping gene biotypes:

- 74% protein-coding
- 22% lncRNA
- 3% pseudogene types
- 1% others

They include:

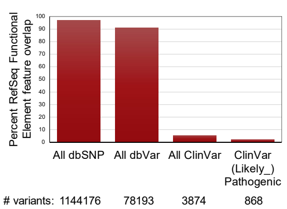
- 797 RefSeqGene (RSG) genes
- 122 Locus Reference Genomic (LRG) genes
- 638 genes used for ClinVar submissions

> ➢ Significant gene overlap is observed
> ➢ Gene regions may not always have exclusive transcript or protein-coding function!

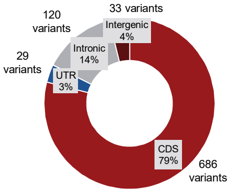
Gene association caution:

- ➢ Gene regulatory elements don't necessarily regulate the genes they overlap or are closest to
- ➢ ***Example:*** the *ZRS* limb enhancer (GeneID:105804841) in an intron of the *LMBR1* gene regulates *SHH* located 1 Mb away!

Feature overlap with variation datasets



# variants:	All dbSNP	All dbVar	All ClinVar	ClinVar (Likely_) Pathogenic
	1144176	78193	3874	868



Locations of ClinVar (Likely_) Pathogenic variants that overlap Functional Elements

Implications for variant interpretation:

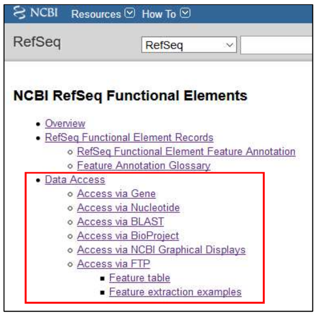
- ➢ Not all disease-associated variation may be due to alterations in protein-coding or transcript function. Mutations in overlapping elements with non-transcript-related function may be disease-causing too!

Also check out NCBI's **gene/transcript-related RefSeq** annotation, including the new **Matched Annotation from NCBI and EMBL-EBI** (MANE) dataset: www.ncbi.nlm.nih.gov/refseq/MANE/

Data access

3

Find multiple ways to access our data on our website: ncbi.nlm.nih.gov/refseq/functionalelements/



NCBI Resources How To

RefSeq | RefSeq

NCBI RefSeq Functional Elements

- Overview
- RefSeq Functional Element Records
  - RefSeq Functional Element Feature Annotation
  - Feature Annotation Glossary
- Data Access
  - Access via Gene
  - Access via Nucleotide
  - Access via BLAST
  - Access via BioProject
  - Access via NCBI Graphical Displays
  - Access via FTP
    - Feature table
    - Feature extraction examples

RefSeq records

4

- Genomic accessions with **NG_ prefixes**
- Associated with NCBI **BioProject accession PRJNA343958**
- International Nucleotide Sequence Database Collaboration (INSDC) feature annotation
- Additional controlled vocabularies from the Sequence Ontology
- Experimental evidence qualifiers with:
  - the type of **evidence** (Evidence & Conclusion Ontology codes)
  - links to **publications** in PubMed
- Additional qualifiers with further details from the literature, including */function*, */note* or */bound_moiety* qualifiers

```
     regulatory      867..22391
                     /regulatory_class="locus_control_region"
                     /experiment="EXISTENCE:transgenic organism evidence
                     [ECO:0001131][PMID:3690667]"
                     /note="21.5 kb ClaI-BglII fragment from -1 kb to -22.5 kb;
                     includes 5'HS1-5'HS5"
                     /function="regulates developmental expression of the
                     beta-globin genes"
                     /db_xref="GeneID:109580095"
     regulatory      3114..5057
                     /regulatory_class="matrix_attachment_region"
                     /experiment="EXISTENCE:fractionation evidence
                     [ECO:0000100][PMID:3208739, PMID:2559410]"
                     /note="2 kb AvaII pGSE228 fragment c SAR"
                     /function="associates with the nuclear scaffold"
                     /db_xref="GeneID:109580095"
     regulatory      5274..6209
                     /regulatory_class="enhancer"
                     /experiment="EXISTENCE:reporter gene assay evidence
                     [ECO:0000049][PMID:20231293, PMID:17548470]"
                     /note="5'HS1 or HS1-3' enhancer fragment"
                     /function="enhancer in K562 erythroleukemia cells"
                     /db_xref="GeneID:109580095"
```

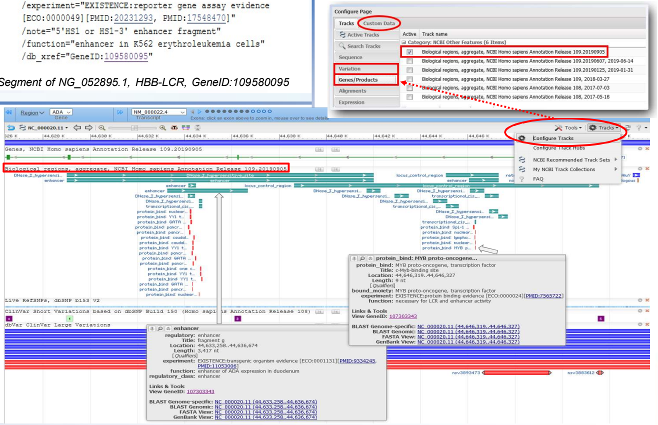
Example: *Segment of NG_052895.1, HBB-LCR, GeneID:109580095*

NCBI genome browser view

5

View RefSeq Functional Elements **feature annotation** in an NCBI genome browser, including the Genome Data Viewer (ncbi.nlm.nih.gov/genome/gdv/) or the Variation Viewer (ncbi.nlm.nih.gov/variation/view/):

- Configure to display the **'Biological regions'** track (see top right of image)
- Load NCBI 'Genes' or variation tracks (e.g. dbSNP, ClinVar, dbVar), configure a track hub of interest, or use the 'Custom Data' tab in the configuration interface to **view alongside other data of choice**
- Mouse-over a Functional Element feature to view the **associated tooltip** with relevant functional data, including links to publications and sequences



Configure Page

Tracks | Custom Data

Active Tracks; Search Tracks; Sequence; Variation; Genes/Products; Alignments; Expression

Active	Track name
	⊟ Category: NCBI Other Features (6 Items)
☑	Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109.20190905
☐	Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109.20190607, 2019-06-14
☐	Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109.20190125, 2019-01-31
☐	Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109, 2018-03-27
☐	Biological regions, aggregate, NCBI Homo sapiens Annotation Release 108, 2017-07-03
☐	Biological regions, aggregate, NCBI Homo sapiens Annotation Release 108, 2017-05-18

Region: ADA (Gene) | NM_000022.4 (Transcript) | Exons: click an exon above to zoom in, mouse over to see details

NC_000020.11 | Tools | Tracks: Configure Tracks; Configure Track Hubs; NCBI Recommended Track Sets; My NCBI Track Collections; FAQ

Genes, NCBI Homo sapiens Annotation Release 109.20190905

Biological regions, aggregate, NCBI Homo sapiens Annotation Release 109.20190905

Live RefSNPs, dbSNP b153 v2

ClinVar Short Variations based on dbSNP Build 150 (Homo sapiens Annotation Release 108)

dbVar ClinVar Large Variations

protein_bind: MYB proto-oncogene...
protein_bind: MYB proto-oncogene, transcription factor
Title: c-Myb-binding site
Location: 44,646,319..44,646,327
Length: 9 nt
[Qualifiers]
bound_moiety: MYB proto-oncogene, transcription factor
experiment: EXISTENCE:protein binding evidence [ECO:0000024][PMID:7565722]
function: necessary for LCR and enhancer activity
Links & Tools
View GeneID: 107303343
BLAST Genome-specific: NC_000020.11 (44,646,319..44,646,327)
BLAST Genomic: NC_000020.11 (44,646,319..44,646,327)
FASTA View: NC_000020.11 (44,646,319..44,646,327)
GenBank View: NC_000020.11 (44,646,319..44,646,327)

enhancer
regulatory: enhancer
Title: fragment g
Location: 44,633,258..44,636,674
Length: 3,417 nt
[Qualifiers]
experiment: EXISTENCE:transgenic organism evidence [ECO:0001131][PMID:9334245, PMID:11053006]
function: enhancer of ADA expression in duodenum
regulatory_class: enhancer
Links & Tools
View GeneID: 107303343
BLAST Genome-specific: NC_000020.11 (44,633,258..44,636,674)
BLAST Genomic: NC_000020.11 (44,633,258..44,636,674)
FASTA View: NC_000020.11 (44,633,258..44,636,674)
GenBank View: NC_000020.11 (44,633,258..44,636,674)

Example: *Genome Data Viewer browser image showing LOC107303343 feature annotation overlapping the ADA gene and variation data on chromosome 20. Tooltips are displayed for two select features.*

U.S. National Library of Medicine
National Center for Biotechnology Information
8600 Rockville Pike, Bethesda MD, 20894 USA

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine

***RefSeq Curators:** C.M. Farrell, T. Goldfarb, S.H. Rangwala
RefSeq Developers: T.D. Murphy, A. Astashyn, O. Ermolaeva, V. Kodali, C. Wallin

www.ncbi.nlm.nih.gov/refseq/functionalelements/